DESCRIPTIVE STATISTICS

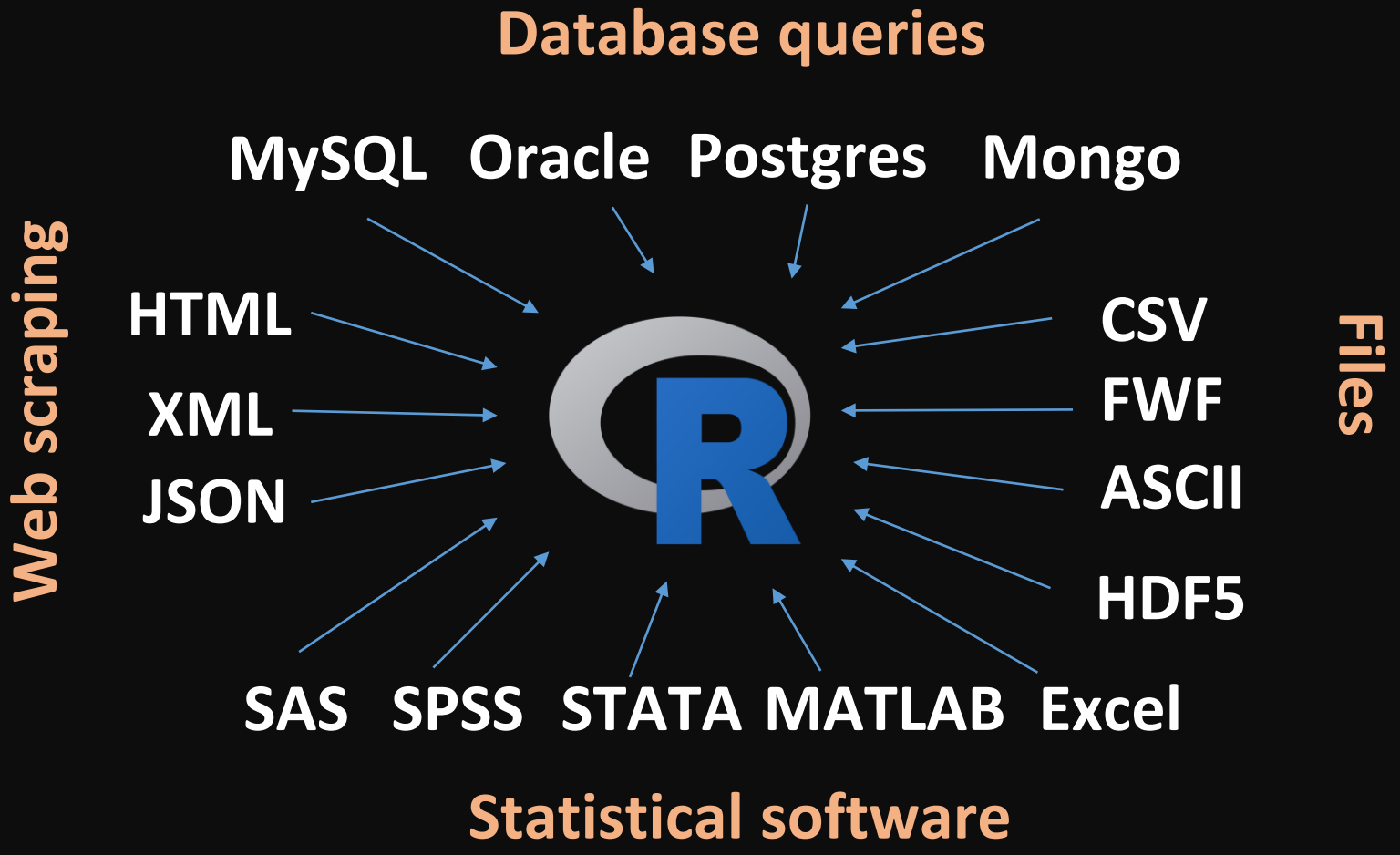
Zack Kertcher

Statistics for Management
Fall 2016

Plan for today

1. Loading data (recap)
2. Descriptive statistics (univariate)

LOADING DATA RECAP


Database queries
MySQL
Oracle
Postgres
Mongo
Web scraping
HTML
XML
JSON
CSV
FWF
ASCII
HDF5
Files
SAS
SPSS
STATA
MATLAB
Excel
Statistical software
Compressed files
Many other…

Reading data from a file

Function	Used for reading
readLines	raw text files
read.csv	csv files separated by a ","
read.csv2	csv files separated by a ";"
read.delim	files separated by "\t"
fead.fwf	fixed format files
read.table	all the above plus more

The read.csv function

```
df <- read.csv(file.choose(), header=T, stringsAsFactors=F)
```


Data frame name

Or specify file location, e.g.
`"c:/downloads/somefile.csv"`

Inspect data

```
dim|ncol|nrow()
```
← df size

```
head|tail(df,n=10)
```
← view portions of df

```
names|colnames|rownames()
```
← column names

```
View()
```
← view in GUI editor

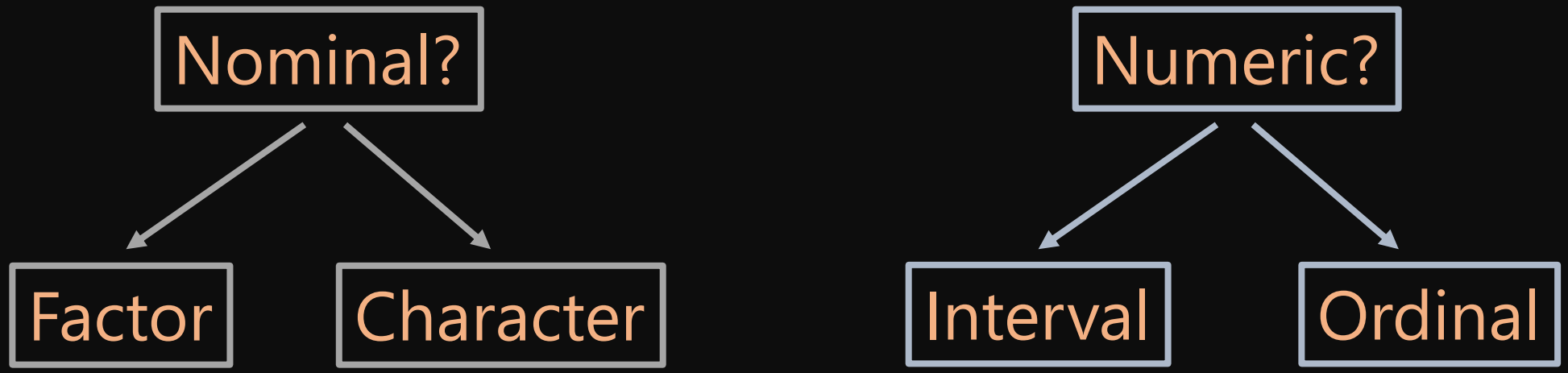
```
fix()
```
← edit values in GUI editor

- ✔ Download the two datasets found under today's folder in Blackboard.
- ✔ Use **cars** data frame to read cars.csv. We will work on it throughout class.
- ✔ Read realestate.csv as **realestate** data frame. We will work on it throughout our exercises today.
- ✔ What are the dimensions of the **realestate** data?
- ✔ How many numeric variables are in the **realestate** data and how many are factor?
- ✔ Does any of these variables need to be converted into a different type? (e.g., numeric to factor, or factor to numeric)

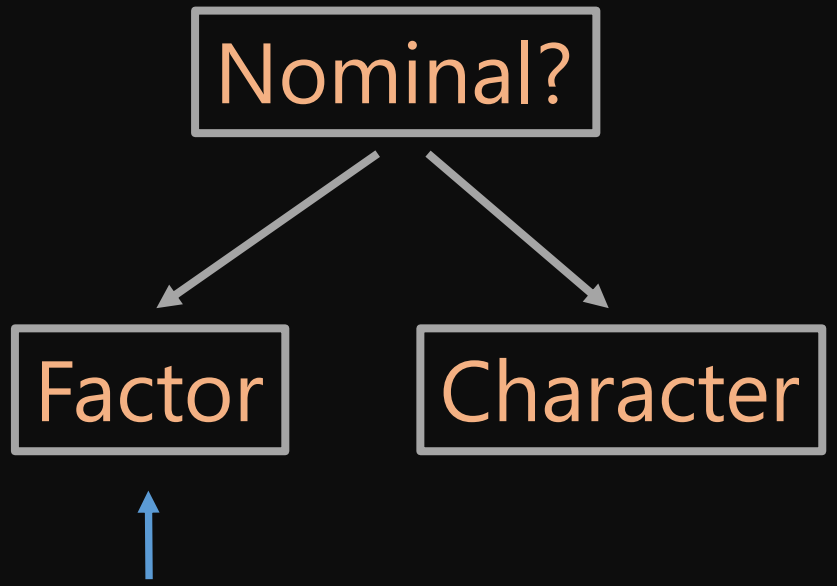
DESCRIPTIVE STATISTICS (UNIVARIATE)

What are the ways to describe a variable?

Identify the variable type



Nominal variable

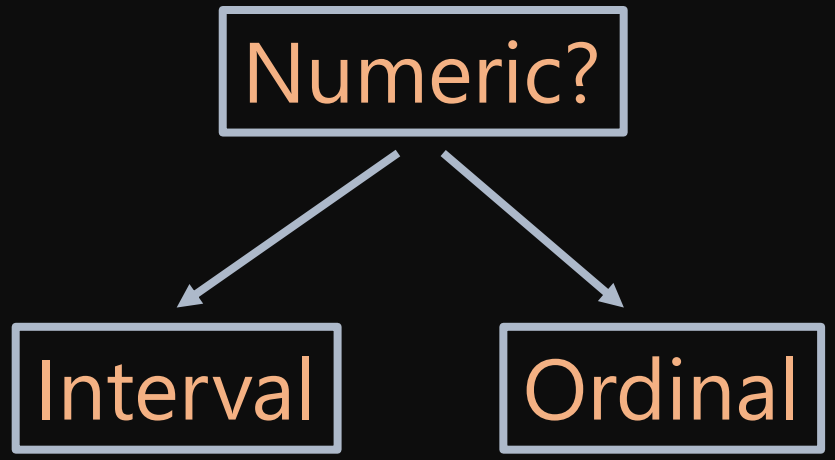


Stats: table( )

Vis: barplot( ) |pie() |plotrix::pie3D

- ✔ Find the distribution of houses based on preferable location (**prefarea**), then convert it to percentages.
- ✔ Plot the proportion of houses with/without **driveway**, using a barplot. Use orange and darkgreen respectively to color the bars. Title the plot: Distribution of Driveway in Real Estate Market.
- ✔ Now find the proportion of **recroom** in houses that cost less than or equal to $350k and compare to houses priced over $350k.

Numeric variable

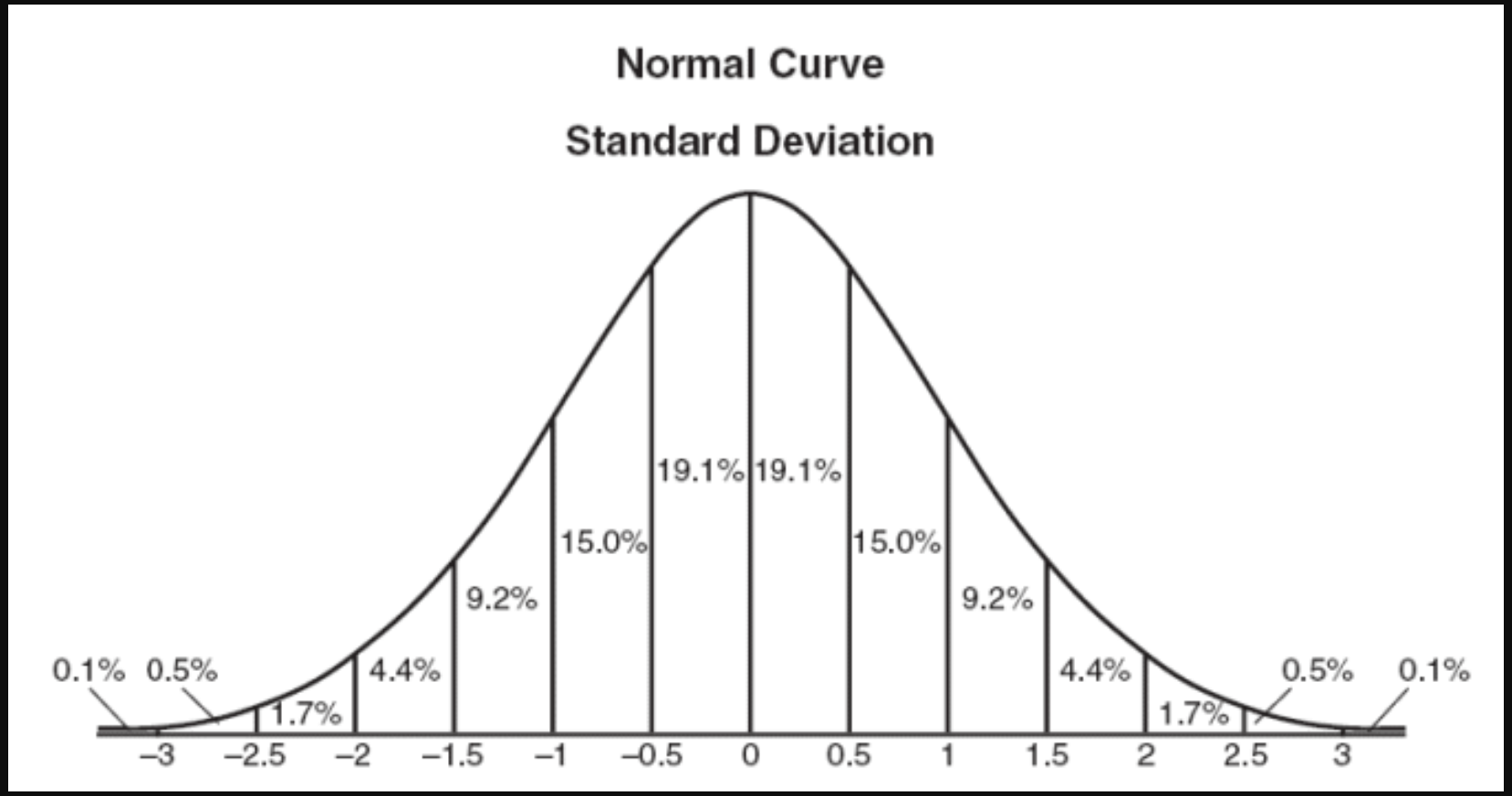


Stats: `mean|median|mode|range|min|max|quantile…()`

Vis: `dotchart|hist|boxplot|density|rug()`

What do these functions measure?

Distribution, especially central tendency



More on distributions next class

- ✔ What is the **price** range in the **realestate** data frame?
- ✔ What is the difference between the mean and median **price**?
- ✔ What are the 40th and 97th percentile of house prices of houses over $400k?

- ✔ Which of the numeric variables in **realestate (lotsize, price, bedrooms)** follows a normal distribution? Explain.

- ✔ Hint: Use variance, standard deviation, MAD, skewness, and kurtosis to reach a conclusion.